Making your agency's sites more accessible to search engine users

Implementing the Sitemap protocol

Agenda

Common barriers to search engine crawling

How citizens access their government online

Implementing the Sitemap protocol

Google Webmaster Tools

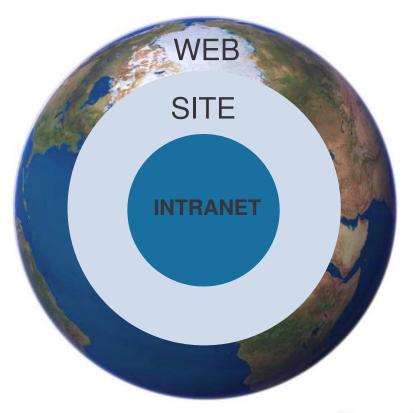
Success stories

Q&A



Clarifications

- Public content only
- Non-proprietary
- No direct cost—nothing for sale
- Web search, not site search (e.g. Google Search Appliance)
- No security risk





Sitemaps.org

An open, industry standard for web search engine crawling



What are Sitemaps?

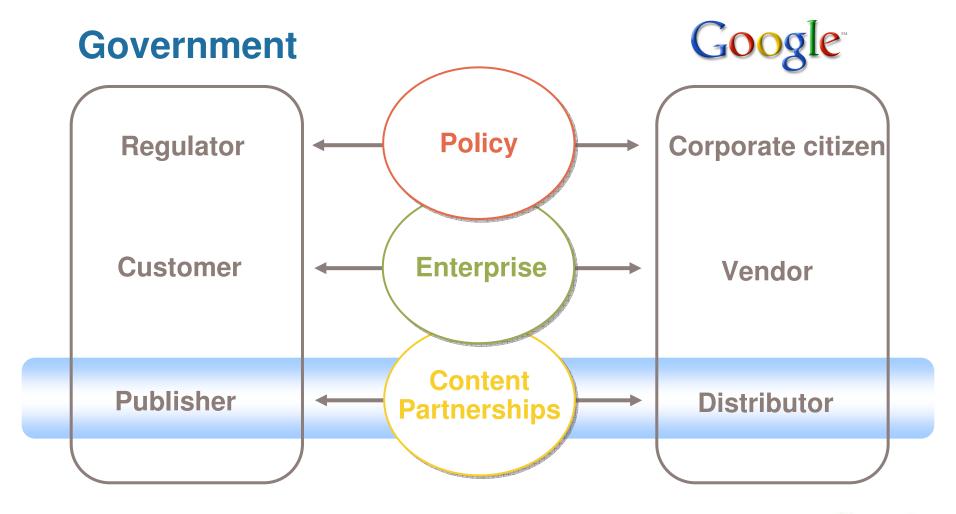
Sitemaps are an easy way for webmasters to inform search engines about pages on their sites that are available for crawling. In its simplest form, a Sitemap is an XML file that lists URLs for a site along with additional metadata about each URL (when it was last updated, how often it usually changes, and how important it is, relative to other URLs in the site) so that search engines can more intelligently crawl the site.

Web crawlers usually discover pages from links within the site and from other sites. Sitemaps supplement this data to allow crawlers that support Sitemaps to pick up all URLs in the Sitemap and learn about those URLs using the associated metadata. Using the Sitemap protocol does not guarantee that web pages are included in search engines, but provides hints for web crawlers to do a better job of crawling your site.

Sitemap 0.90 is offered under the terms of the Attribution-ShareAlike Creative Commons License and has wide adoption, including support from Google, Yahoo!, and Microsoft.



Google's relationships with government





Web search vs. site search

Supporting the two levels of search



Site Search

All of the open and accessible deep web	Search scope	A segment of your public site's content
Citizens and professionals	User	Professionals and citizens
Googlebot's crawling intervals	Freshness	Customizable
Limited by robots.txt, dynamic content.	Crawling	Limited by server capacity and cost
High-level stats	Reporting tools	More detailed, all facets
Free	Cost	Varies



The majority of citizens access government through web search engines

National Institutes of Health (nih.gov)

 More than 70% of unique users in July 2006 were referred by search engines (Google, Yahoo, MSN, AOL, Ask)



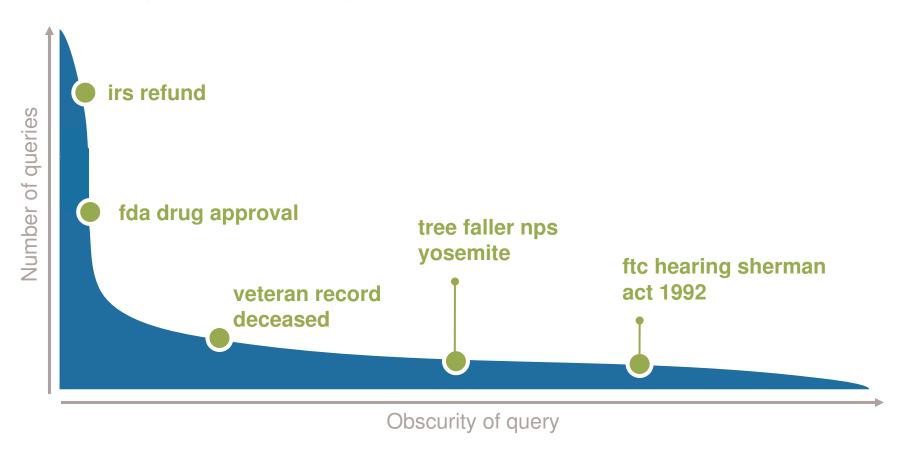
Only 4% of unique users came directly to nih.gov sites



Source: ComScore, 2006

And citizens expect to find everything through search

The long tail of federal government information





Not all information is created equal

The value of government content – a pillar of the web

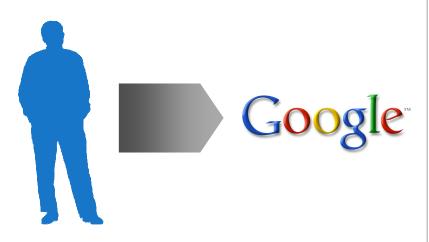
Government

Untrustworthy

Trustworthy



Search engines are the point of departure, government sites are the destination



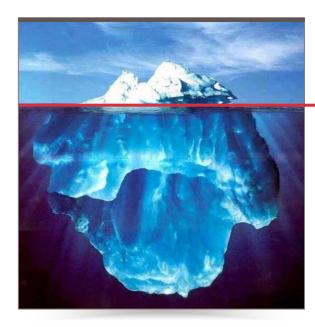




Barriers to web search engine crawling

What can make a site effectively invisible to search engine users

- Content "hidden" behind search forms
- Non-HTML links
- Outdated robots.txt crawling restrictions
- Server errors (crawler times out when fetching content)
- Orphaned URLs
- Rich media: audio, video
- Premium content



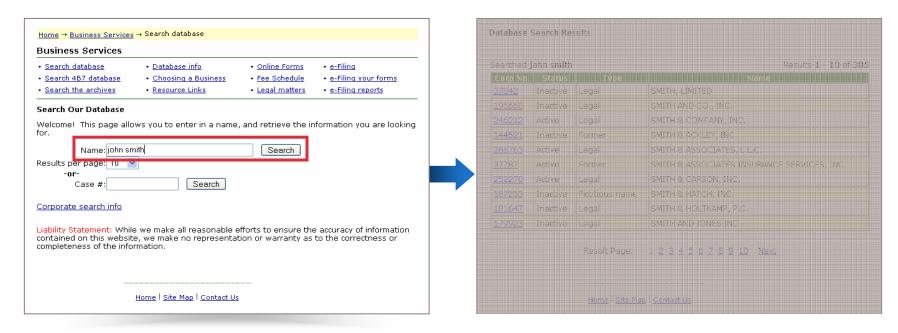
WEB Searchable

DEEP WEBNot searchable



Crawlers cannot navigate search forms

When crawled



Search results are invisible



The solution: Sitemaps

The Sitemap protocol enables a web publisher to proactively manage search engine crawling







"The launch of Sitemaps is significant because it allows for a single, easy way for websites to provide content and metadata to search engines"

—Tim Mayer, Senior Director of Product Management, Yahoo Search

"We are 100% behind this protocol - this kind of collaboration will help improve the search experience for all of our customers"

—Ken Moss, General Manager, Live Search

- Sitemap protocol developed by Google in June 2005 and released under **Creative Commons License**
- Adopted as an industry standard in November 2006: www.sitemaps.org



Navigational sitemap

A browse index or sitemap enables a user to navigate throughout a site

SITE INDEX

To view or print the PDF content on this page, download the free Adobe® Acrobat® Reader®.

NEWS

Treasury Deputy Secretary Kimmitt Travels to Asia this week to Discuss Compact with Iraq

KEY TOPICS

General Interest
Law Enforcement
International
Taxes
Financial Markets
Currency & Coins
Small Business
Accounting & Budget
Technology

PRESS ROOM

Public Schedule

OFFICES

Office of Domestic Finance
Debt Management
Advanced Counterfeit Deterrence
Office of Financial Institutions
Federal Financing Bank
Financial Institutions
Financial Markets
Fiscal Service

Office of Economic Policy Working Papers Total Taxable Resources

Office of Foreign Assets Control
Executive Order 13324
National Money Laundering Strategy
Executive Office for Asset Forfeiture

Terrorism and Financial Intelligence



Sitemaps for search engines

- HTML
- Simple text
- XML

15



Simple text sitemap

A comprehensive list of URLs

http://www.firstgov.gov/index.shtml

http://www.firstgov.gov/About.shtml

http://www.firstgov.gov/Citizen/Services/Address_Changes.shtml

http://www.firstgov.gov/Topics/Parents_Adoptive.shtml

http://www.firstgov.gov/Government/State_Local/Ag_Environment.shtml

http://www.firstgov.gov/Citizen/Topics/Environment_Agriculture/Agriculture.shtml

http://www.firstgov.gov/Citizen/Facts/Facts_Agriculture.shtml

http://www.firstgov.gov/Agencies/Federal/Executive/Agriculture.shtml



XML sitemap

- A comprehensive list of URLs in XML
- Tagged with each URL's location, last modification, change frequency and priority

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.google.com/schemas/sitemap/0.84">
   <ur1>
      <loc>http://www.example.com/</loc>
      <lastmod>2005-01-01</lastmod>
      <changefreq>monthly</changefreq>
      <priority>0.8</priority>
   </ur1>
   <ur1>
      <\loc>http://www.example.com/catalog?item=12&amp;desc=vacation hawaii</loc>
      <changefreq>weekly</changefreq>
   </ur1>
   <ur1>
      <loc>http://www.example.com/catalog?item=73&amp;desc=vacation new zealand</loc>
      <lastmod>2004-12-23</lastmod>
      <changefreq>weekly</changefreq>
   </ur1>
   <ur1>
      <loc>>http://www.example.com/catalog?item=74&amp;desc=vacation newfoundland</loc>
      <lastmod>2004-12-23T18:00:15+00:00</lastmod>
      <priority>0.3</priority>
  </ur1>
   <ur1>
      <loc>http://www.example.com/catalog?item=83&amp;desc=vacation usa</loc>
      <lastmod>2004-11-23</lastmod>
   </ur1>
</urlset>
```



Introducing Google Webmaster Tools

Free resources and tools to help you implement sitemaps and improve your sites' visibility in Google search results



Welcome to your one-stop shop for comprehensive info about how Google crawls and indexes websites. You can learn here how to ensure that your site is easily crawled and indexed and access tools that will enable you to diagnose crawling issues, study statistics on how your site is doing in our index, and tell us how you'd like your site to be crawled and indexed.



Site status wizard

Find out whether your site is currently being indexed by Google.



Webmaster tools (including Sitemaps)

Statistics, diagnostics and management of Google's crawling and indexing of your website, including Sitemap submission and reporting.



Submit your content to Google

Learn about submitting content for Google properties such as Google Base and Google Book Search.



Google's blog for webmasters

The latest news and info on how Google crawls and indexes websites.



Google's discussion group for webmasters

Talk with your fellow webmasters and share your feedback with us.



Webmaster help center

See answers to frequently asked questions about crawling, indexing, ranking and other webmaster issues.

Learn more at: http://www.google.com/sitemapsgov



Implementing the Sitemap protocol



Step 1: Login to Webmaster Tools with your Google Account



Google's webmaster tools provide you with a free and easy way to make your site more Google-friendly. Using our tools, you can:

Get Google's view of your website, and diagnose potential problems.

See how Google crawls and indexes your site and learn about specific problems we're having accessing it.

See how your site is performing.

Learn which queries drive traffic to your site, and see exactly how users arrive there.



Share info with us to help us crawl your site better.

Tell us about your pages: which ones are most important to you and how often they change. You can also let us know how you would like the URLs we index to appear.

Get started today -- it's free!

Simply log in with your Google Account and add your site URL to get started. It's an easy and free way to have a more interactive experience with Google.

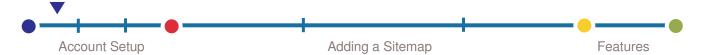




Login

Learn more about Google webmaster tools:

- About Google webmaster tools
- Google webmaster central
- Webmaster help center
- Google webmaster discussion group





Step 2: Add a Site to verify ownership



mygovsite@gmail.com | My Account | Help | Sign out

Google webmaster tools are an easy way for you to submit all your URLs to the Google index and get detailed reports about the visibility of your pages on Google. To get started, simply add the URL of your site. You'll start to see information about your site right away.



⊕ Tools

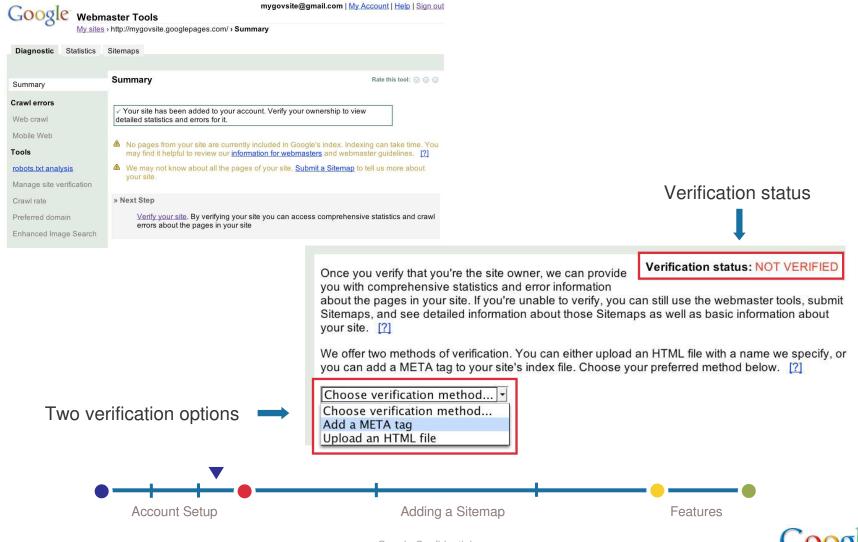
Read more about Google webmaster tools:

Learn more about the Google webmaster tools program
Learn more about Google Mobile Sitemaps



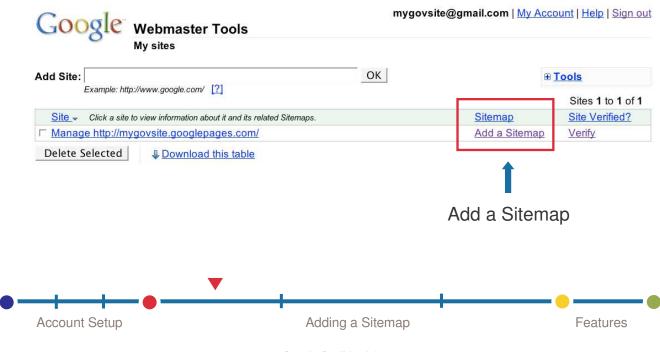


Step 3: Verify your site



Step 4: Add a sitemap

- A. Create a sitemap with Google's Sitemap Generator or any third-party tool
 - Use any available 3rd party tool (http://code.google.com/sm_thirdparty.html)
 - For custom, dynamic environments, you may need to rely on internal scripts to generate a list of urls





Step 4: Add a sitemap

A. Create a sitemap with Google's Sitemap Generator or any third-party tool

Which sitemap is best for your site?

	Simple Text	XML
pros	+ Easy to create + Acceptable format	+ Provides detailed information for smarter, efficient crawling + Tags are optional
cons	- No gains in efficiency	- Entails additional steps



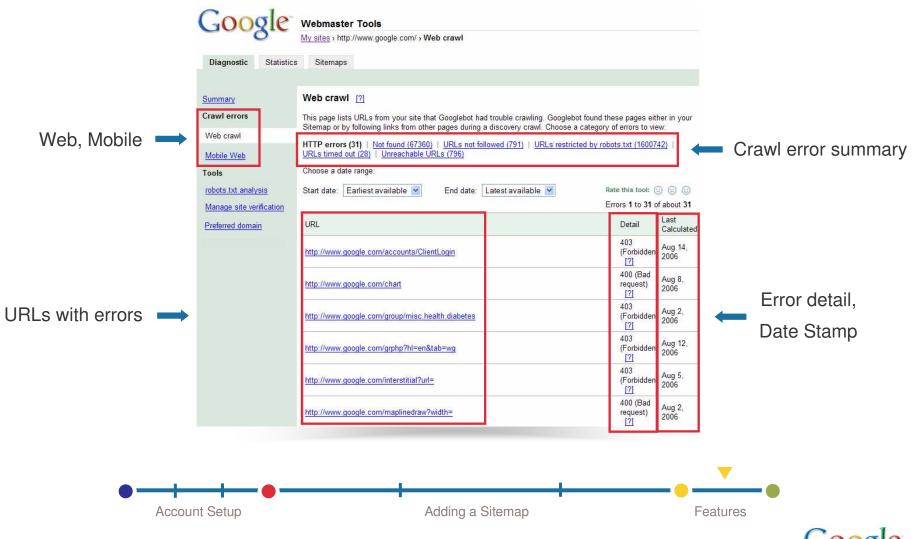


Step 4: Add a sitemap

- B. Upload the sitemap file to your site
- C. Add the sitemap URL to your account
 - Add at the highest level in your website directory structure that you want crawled
 - See the status of the sites and sitemaps in your account



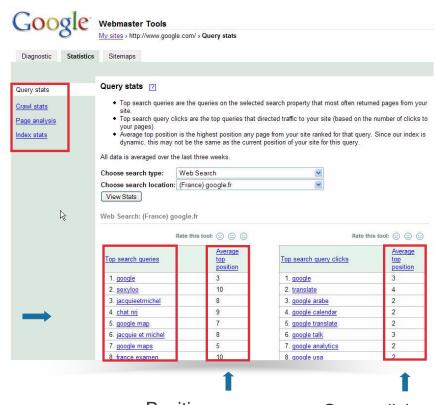
Features: Crawl Errors Show problematic pages



Features: Query Stats identifies popular queries

- See your top 20 search queries and search query clicks
- Top position shows you where your pages were listed per search query
- Easily export a report with CSV download feature

Search queries = impressions in search results



Position per query in results

Query clicks = traffic





Overview of features: More information and statistics

- Crawl Errors → shows you which pages were problematic

- Robots.txt

 helps to improve your coverage
- Page Analysis → shows how Google sees your pages





Some questions to consider

- Publishing system:
 - What database applications (Oracle, SQL Server, flat files, etc) do you operate?
- System management:
 - Can you download and install third-party tools on your web server?
- URL structure:
 - Can you list and explain a few combinations of how your site URL is constructed?



Success stories



PlainLanguage.gov success story

- Plain Language Information and Action Network (PLAIN), a federal inter-agency volunteer working group that encourages clarity in government communication to the public through PlainLanguage.gov
- Before-and-after examples of government documents served dynamically, thus uncrawlable





PlainLanguage.gov success story



- Web manager successfully implemented sitemap in ~8 hours, using available resources and through trial and error
- As new examples are added to the database, the sitemap is regenerated and submitted



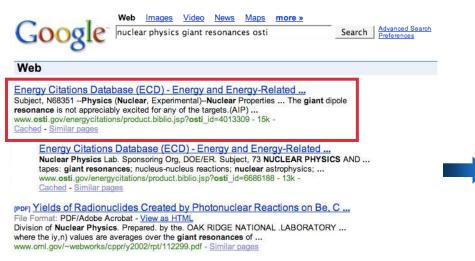
OSTI success story

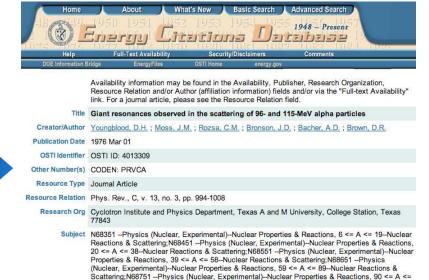
- Department of Energy agency that "makes R&D findings available and useful, so that science and technological creativity can advance"
- Web manager submitted sitemaps for Energy Citations and Information Bridge services, opening 2.3M bibliographic records and full-text documents to crawling
- Sitemap standard assures web search engines have "a complete picture" of information in OSTI services





OSTI success story





- Benefits include better representation in search results and reduced load on servers (by limiting duplicate crawling)
- First implementation completed in 16 staff hours -- can now be easily replicated across web search engines



NCES success story

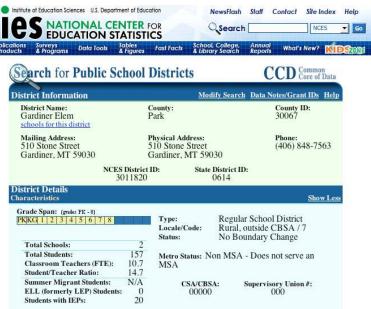
- Department of Education agency that provides statistical information about districts, schools, and other educational facilities
- Using freely available tools, web manager submitted sitemaps to open five dynamic databases to crawling, adding 180K URLs





NCES success story





- Now surfacing tens of thousands of potential web search hits with links to NCES services
- Helping to ensure citizen users gain access to the latest data from the original source



Federal Sitemaps

An initiative to help federal agencies make their sites more accessible to search engine users

FederalSitemaps

WikiHomePage | RecentChanges | Page Index



Login (create account)

Federal Sitemaps

Upcoming and Past Events (3E66

oming and Past Events (3EBG)

- The <u>Sitemap protocol</u> is an open, XML-based standard for managing search engine crawling. The protocol provides website owners a means of communicating to search engines the location, priority, change frequency, and last modification date of all pages on a website or web-accessible database, which can ensure complete and efficient crawling of the site's contents. (3CFM)
- The Sitemap protocol was introduced by Google in June 2005 under a Creative Commons License and was adopted in November 2006 as an industry standard by Google, Microsoft and Yahoo. (3CFN)
 - SearchEngineWatch Search Engines Unite On Unified Sitemaps System (3CQI)

Your Visited Pages

FederalSitemaps

View Backlinks

Search

http://tinyurl.com/3byhy7



Relevant legislation and OMB policy

- The Sitemap protocol supports the E-Government Act of 2002 requirements to:
 - "Organize and categorize information intended for public access and ensure it is searchable across agencies...[using] sophisticated Internet search functions (including their crawl and index mechanisms)..."
 - "...publish your information directly to the Internet...expos[ing] information to freely available and other search functions [that] adequately [organize] and [categorize] your information."
 - "...[When] disseminating significant information dissemination products, advance preparation, such as using formal information models, may be necessary to ensure effective interchange or dissemination. This procedure is needed when freely available and other search functions do not adequately organize and categorize your information."
- The Sitemap protocol also supports the Federal Enterprise Architecture's Data Reference Model 2.0 requirements to:
 - "Identify how information and data are created, maintained, accessed, and used...[and] Define data and describe relationships among data elements used in the agency's information systems."
 - * Source: Federal Sitemaps (http://colab.cim3.net/cgi-bin/wiki.pl?FederalSitemaps), Semantic Interoperability (SICoP) and XML (xmlCoP) Community of Practice.



Next steps for web managers: Prepare

- Audit your agency's sites to identify uncrawlable elements
 - Google can provide support with analysis: sitemappartners@google.com
 - Sample sitemapping target list:
 http://spreadsheets.google.com/pub?key=pUb62ZKHnzgqEoGF4L
 Ff3Gw
- Get trained:
 - Attend Google's webinar on technical steps to implementing sitemaps: Thursdays, 3:00 EST
 - Or arrange dedicated webinar for your agency



Next steps for web managers: Implement

- Sign up at www.google.com/sitemapsgov:
 - Verify your sites' ownership
 - Produce and upload sitemaps
- Get answers:
 - At Webmaster Central: www.google.com/webmasters
 - Or directly: sitemap-partners@google.com
- Track your progress



Making your agency's sites more accessible

- Implementing sitemaps can enhance, but does not replace, a web search engine's crawling
- It does not guarantee inclusion, but helps to provide users more information and fresher results
 - The Sitemap protocol is an open, industry standard
 - Ensures all your agency's public information and services are discoverable by all potential users
 - Also accelerates the inclusion of new information in search results

- Makes web search engine crawling more efficient, reducing demands on servers
- Most sitemaping tools are free and can be easy to implement
- Can be readily replicated across web search engines

